AI & The Cyber Security Frontier

SIGS Basel, 17th October 2024

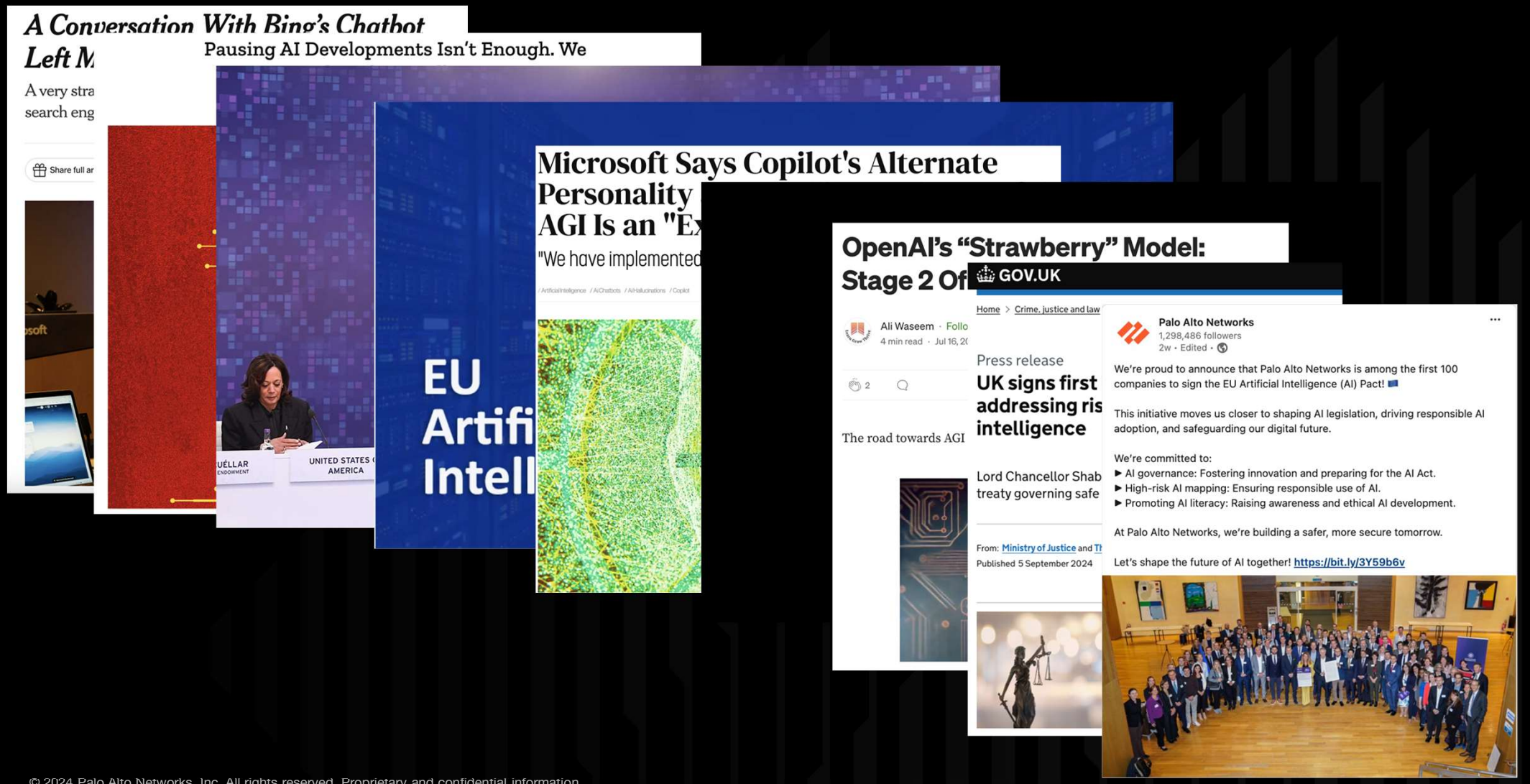
Peter Bury
pbury@paloaltonetworks.com

Topics

- What do you mean, “AI”? … Themes and Perceptions
- AI Secure Usage
- AI in Cybersecurity

What do you mean, “AI”?

2 Years of GenAI



Emerging Themes, Legislation

AI Factions: Accelerationists vs Altruists

US Executive Order (30/10/23) - Safe, Secure, and Trustworthy Development and Use of AI

The Bletchley Declaration (01/11/23) - An overarching commitment to the design, development, deployment and use of AI in a manner that is safe, human-centric, trustworthy and responsible.

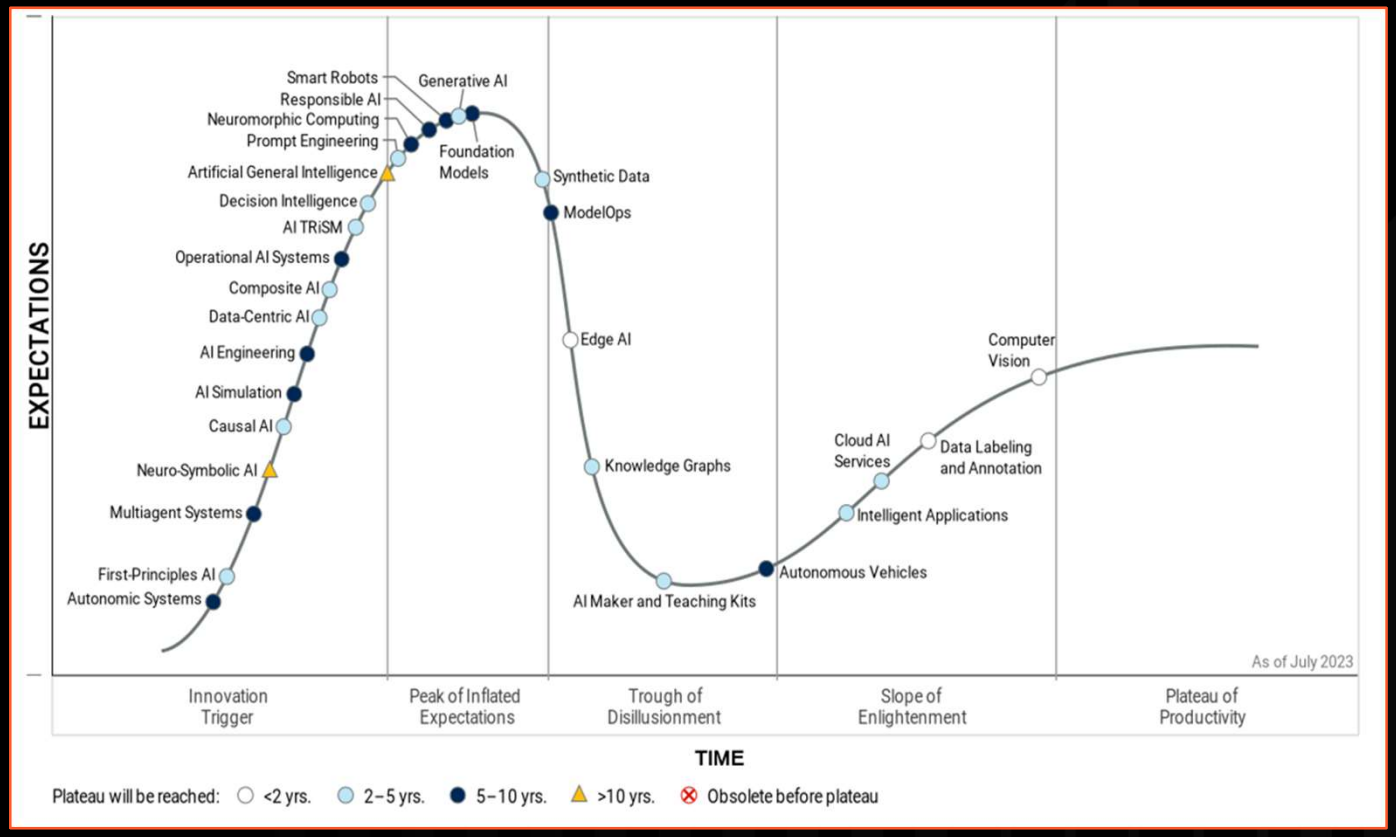
EU AI Act (01/08/2024) Establish a consumer protection-driven approach through a risk-based classification of AI technologies as well as regulating AI more broadly.

SB-1047 (28/08/2024) Safe & Secure Innovation for Frontier Artificial Intelligence Models Act

U.S. AI Safety Institute/NIST (29/08/24) Agreements enabling formal collaboration on AI safety research, testing and evaluation with Anthropic and OpenAI.

https://www.aisafetysummit.gov.uk/
https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Adoption and Integration



Resources - education and talent

The limits of LLMs - Hallucinations?

Transformer Architecture?
(Mamba, Megalodon)

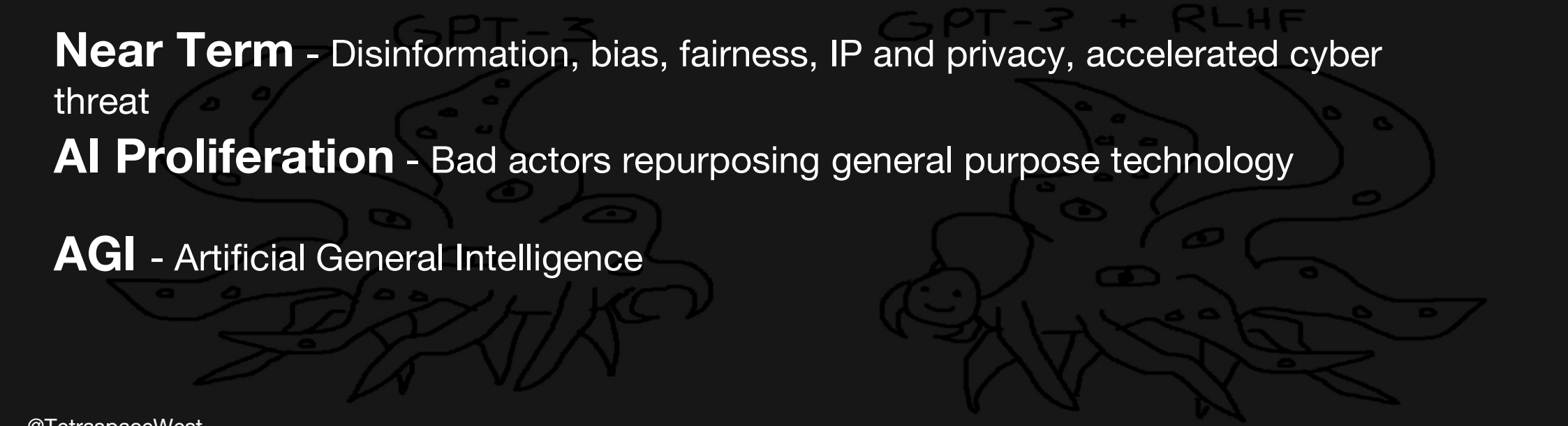
Security and Integrity of data in public systems

Legal and copyright

Many organizations lack the data practices necessary to input data into AI

Demis Hassabis, CEO, Google DeepMind

"I would advocate not moving fast and breaking things"



Near Term - Disinformation, bias, fairness, IP and privacy, accelerated cyber threat

AI Proliferation - Bad actors repurposing general purpose technology

AGI - Artificial General Intelligence

@TetraspaceWest

https://www.lesswrong.com/

https://www.theintrinsicperspective.com/p/the-banality-of-chatgpt

Threats TO AI - Adversarial Attacks

ATLAS™

The ATLAS Matrix below shows the general progression of attack tactics as column headers from left to right, with attack techniques organized below each tactic. & indicates a tactic or technique directly adapted from from ATT&CK. Click on the blue links to learn more about each item, or search and view more details about ATLAS tactics and techniques using the links in the top navigation bar.

Reconnaissance &	Resource Development &	Initial Access &	ML Model Access	Execution &	Persistence &	Privilege Escalation &	Defense Evasion &	Credential Access &	Discovery &	Collection &	ML Attack Staging	Exfiltration &	Impact &
5 techniques	7 techniques	6 techniques	4 techniques	3 techniques	3 techniques	3 techniques	3 techniques	1 technique	4 techniques	3 techniques	4 techniques	4 techniques	6 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak		Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access						LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection											Cost Harvesting
	Poison Training Data	Phishing &											External Harms
	Establish Accounts &												

https://arxiv.org/pdf/2310.13828.pdf
https://arxiv.org/abs/2310.13828
https://atlas.mitre.org/
https://github.com/mitre/advmlthreatmatrix

Threats TO AI - Adversarial Attacks … And Human Error

The Register

Exposed Hugging Face API tokens offered full access to Meta's Llama 2

With more than 1,500 tokens exposed, research highlights importance of securing supply chains in AI and ML

Connor Jones — Mon 4 Dec 2023 // 14:00 UTC

The API tokens of tech giants Meta, Microsoft, Google, VMware, and more have been found exposed on Hugging Face, opening them up to potential supply chain attacks.

Researchers at Lasso Security found more than 1,500 exposed API tokens on the open source data science and machine learning platform – which allowed them to gain access to 723 organizations' accounts.

In the vast majority of cases (655), the exposed tokens had write permissions granting the ability to modify files in account repositories. A total of 77 organizations were exposed in this way, including Meta, EleutherAI, and BigScience Workshop - which run the Llama, Pythia, and Bloom projects respectively.

The three companies were contacted by *The Register* for comment but Meta and BigScience Workshop did not not respond at the time of publication, although all of them closed the holes shortly after being notified.

https://arxiv.org/pdf/2310.138
https://arxiv.org/abs/2310.138
https://atlas.mitre.org/
https://github.com/mitre/advmlthreatmatrix

AI and Cybersecurity

Cyber-Enabled vs Cyber-Dependent Crime

Cyber-enabled crimes

Traditional crimes which can be increased in scale or reach by the use of computers, computer networks or other forms of ICT (such as cyber-enabled fraud and data theft).

Cyber-dependent crimes

Crimes that can be committed only through the use of Information and Communications Technology (ICT). That is, where the devices, networks or systems are the targets of the crime, and the target of the crime

Accelerated Attacks

Scaled Attacks

New Vectors

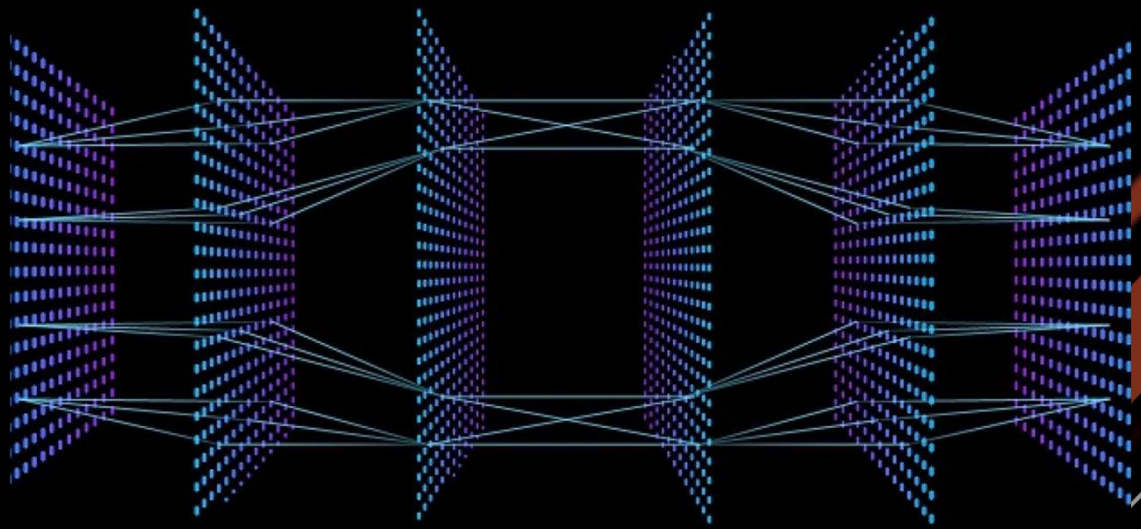
https://www.cps.gov.uk/legal-guidance/cybercrime-prosecution-guidance

It’s All About Data

Data Gravity: Dave McCrory, 2010

More Data -> Smarter & Better Decisions

More Data -> More Complicated Decisions



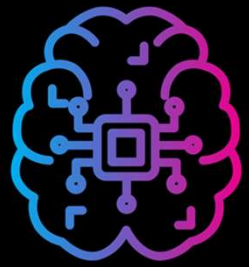
https://datagravitas.com/2010/12/07/data-gravity-in-the-clouds/

Signatures -> ML -> DL

Machine Learning

- Requires real world threat data to be effective
- Data requires significant curation by data scientists
- Threat features must be identified by humans
- Most effective with a small set of threat features

Deep Learning



- A subset of machine learning
- Requires tremendous volumes of real world threat data to be effective
- Data does not require significant curation by data scientists
- Threat features are not identified by humans
- Very effective with large sets of unstructured data

AI in Cybersecurity - Not all AI is created equal

ANI - Artificial Narrow Intelligence

AGI - Artificial General Intelligence

ASI - Artificial Super Intelligence

- **Precision AI**
  - Network
  - Endpoint
  - Identity
  - Cloud
  - Attack Surface
- **AI Copilots**
  - Navigation and Feedback
  - Risk Prioritisation
  - Best Practice Guidance
  - “How-to” answers
  - Case Creation + Resolution
- ***Evolving AI***
  - *Autonomous Agents*
  - *Large Action Models*
  - *Objective Driven AI*
  - *Alternative Architectures*
  - *Retrieval Augmented Generation (RAG)*

AGI - Artificial General Intelligence

ASI - Artificial Super Intelligence

AI in Cybersecurity - Not all AI is created equal

ANI - Artificial Narrow Intelligence

- **Precision AI**
  - Network
  - Endpoint
  - Identity
  - Cloud
  - Attack Surface

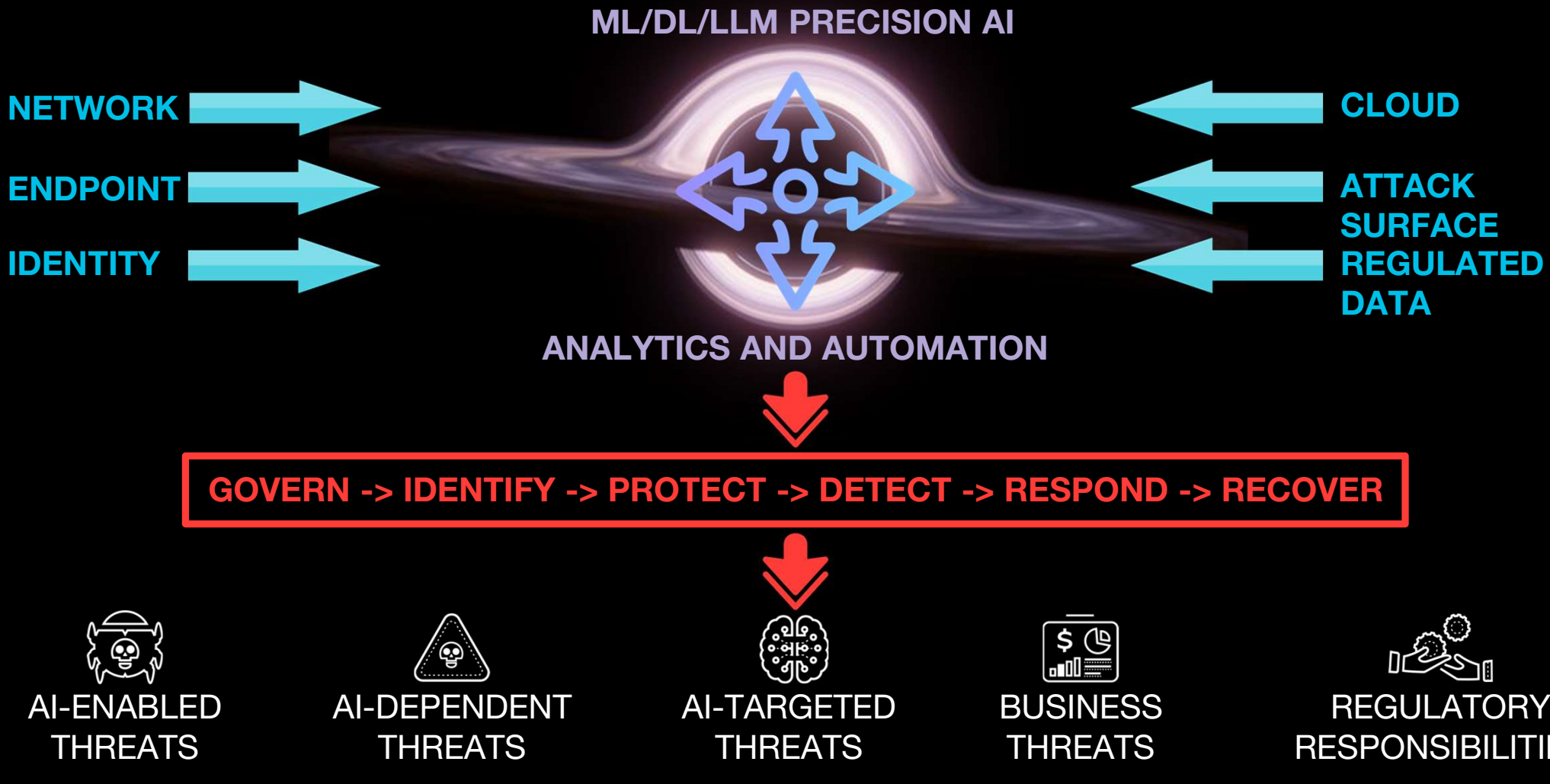
- **AI Copilots**
  - Navigation and Feedback
  - Risk Prioritisation
  - Best Practice Guidance
  - “How-to” answers
  - Case Creation + Resolution

- ***Evolving AI***
  - *Autonomous Agents*
  - *Large Action Models*
  - *Objective Driven AI*
  - *Alternative Architectures*
  - *Human Copilot?*

AGI - Artificial General Intelligence

ASI - Artificial Super Intelligence

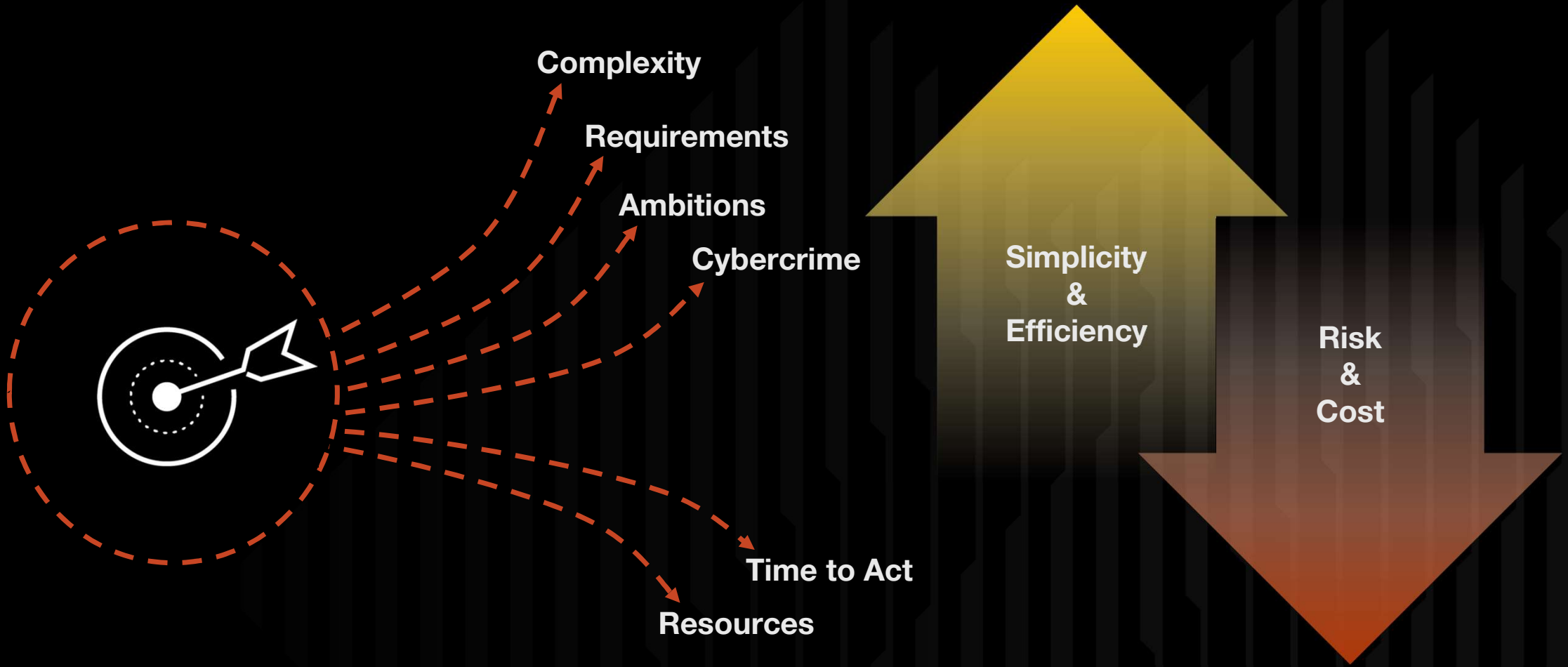
Data + AI for Better CyberSecurity Outcomes



https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.ipd.pdf

Operational Resilience, Policy & AI

The Need for Cyber Resilience



Cyber Resilience Framework

Level 4

- Proactive controls
- Enhanced management
- Highly integrated

Level 3

- Basic controls
- Baseline management
- Some integration

Level 2

- Partially implemented
- Partially managed
- No integration

Level 1

- Not implemented
- Not managed
- No integration

Operational Efficiency

Our Internal Approach to Deploying AI at Palo Alto Networks

Developing an AI Policy

Determining AI Exposure Visibility

Protecting Data Assets

Protecting AI Applications

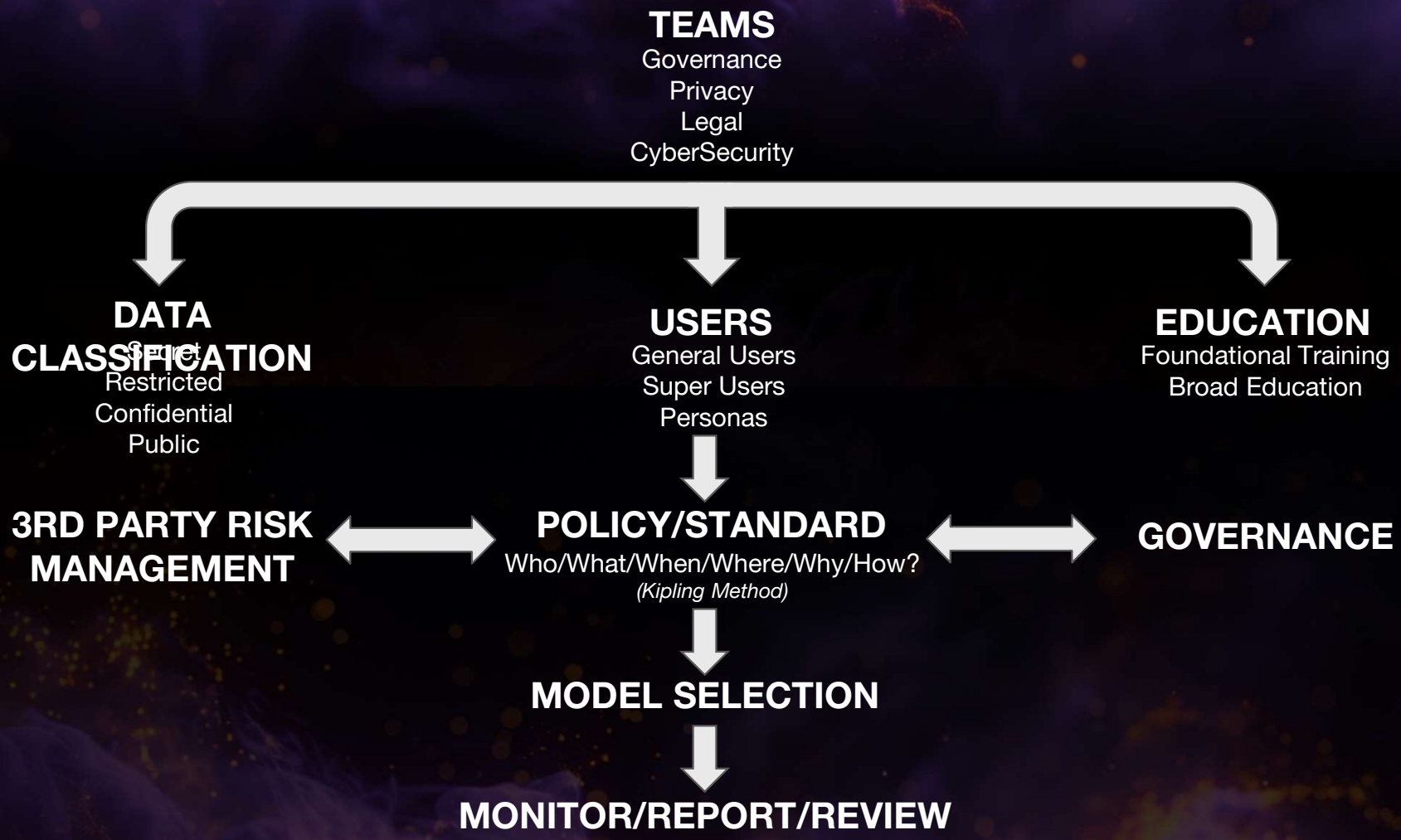
Protecting New HW and Cloud Scale for AI

Identifying and Controlling Access for AI

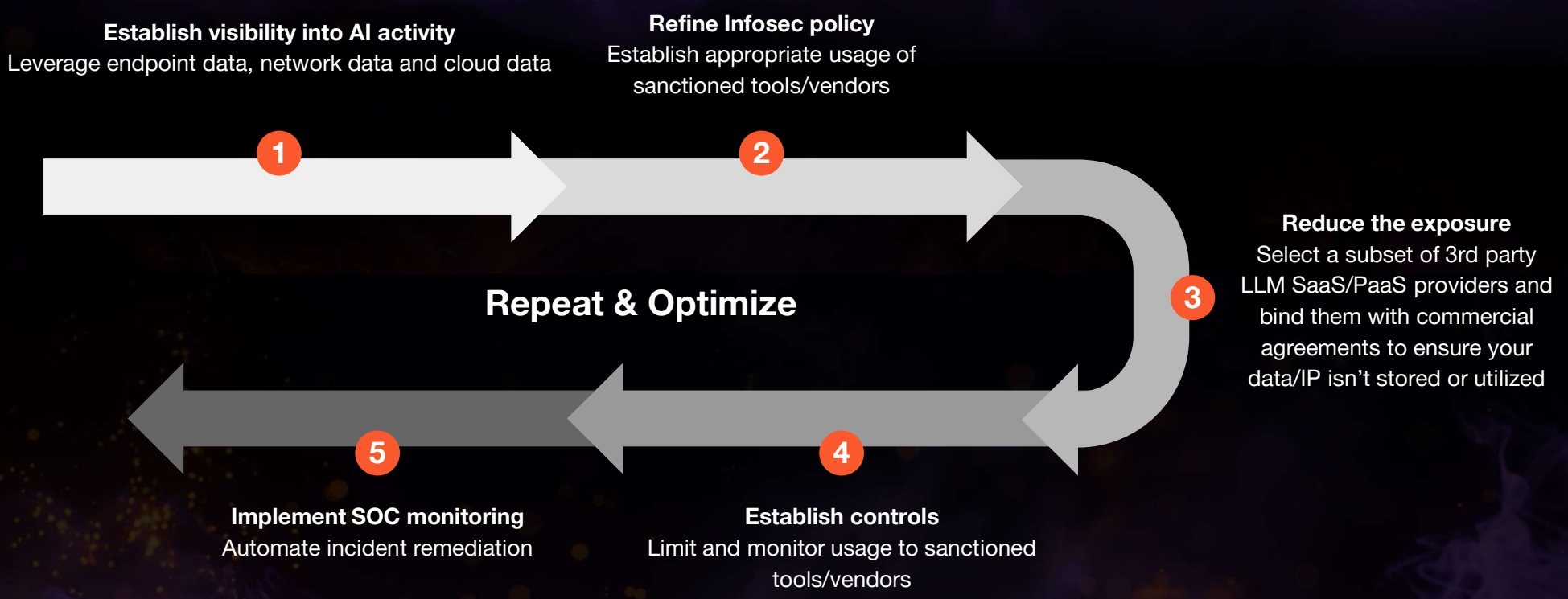
Protecting Models

Securing the AI Supply Chain

Secure AI: Policy & Standards



Secure AI: Visibility and Controls



Thank You

CYBERSECURITY
FOR THE AI ERA